

Salary Prediction Model for Data Related Careers In the United States

Collaborative Project By:

Joseph Choi, Daniel Meier, Serge Nane, Jenny Albrecht



All About The Data

GOAL(S) OF PROJECT:

Create a prediction model of forecasted salary trends in the United States for Data Related Careers.

WHERE IS IT COMING FROM:

[Jobs and Salaries in Data field 2024 \(kaggle.com\)](#)

WHY DID WE CHOOSE THIS:

Personal Interest for All of Us

A decorative pattern at the bottom of the slide consisting of numerous vertical bars of varying heights, rendered in a light teal color against the darker teal background.

Factors to Consider???

- Experience Levels

- Entry Level
- Mid Level
- Senior
- Executive

- Employment Types

- Full Time
- Part Time
- Contract
- Freelance

- Work Setting

- Remote
- In Person
- Hybrid

- Company Size

- Small
- Medium
- Large



Data Related Career Categories

- Data Science and Research
- BI and Visualization
- Data Architecture and Modeling
- Data Analysis
- Data Engineering
- Leadership and Management
- Data Quality and Operations
- Machine Learning and AI
- Cloud and Database
- Data Management and Strategy

Wrangler's Responsibilities


- **Acquire the Data**
- **Clean the Data**
- **Transform the Data**
- **Assess the Data Quality**

COMPLETED DURING THIS STAGE:

- ❏ Imported Libraries and Explore Dataset
- ❏ Standardizing Data within Columns
- ❏ Adjust Column Names
- ❏ Check and Account for Missing Values
- ❏ Ensure Consistency of Data
- ❏ Filter Information for Intended Projection

Science Behind The Results

METHODS USED AND THEORIES ANALYZED



Exploratory Data Analysis

Use to understand the characteristics of each input and extract insights for the feature engineering process.

Built a histogram and box plot for our target variable to analyze its distribution and identify outliers

Built bar charts comparing the average salary with each input categorical feature to understand their distributions and identify any dominant categories that might impact the learning process

Built box plots comparing the salary with each input categorical feature to identify outliers that could affect model training

Built a correlation matrix to comprehend the correlation coefficients between all variables. The output, however, was not helpful as the matrix failed to capture the relationship between our input features (categorical and encoded) and target variable (numerical)



RESULTS OF EDA

The histogram of the salary column displays a bell-shaped curve (close to), indicating a normal distribution. This suggests that linear regression models are advantageous due to their compatibility with normally distributed target variables.

Box plots displayed outliers past the \$300,000 mark, indicating salaries significantly higher than the majority. After a thorough investigation, we concluded that these outliers were legitimate and decided to keep them in our model training process.

Bar charts showed no signs of dominant categories in the dataset, which is essential to avoid bias in the model training process. Each category per feature displayed a balanced progression of increase and decrease, indicating a healthy distribution.

We decided to remove the job title column as it was redundant with the job category column. The job title column contained over 150 unique values that would not have positively contributed to model training.

For encoding, we are using ordinal encoding for work year, experience level, employment type, work setting, and company size to assign numbers to each feature based on order. We are using one-hot encoding for the job category to treat each category independently without assuming any ordinal relationship.

As our correlation matrix failed to capture the correlation coefficients properly, we have decided to pull feature importance information from our linear and non-linear regression models using scikit-learn.

Model Training

MODELS

1. Ridge
2. Lasso
3. Random Forest
4. Gradient Boosting Machines

ITERATIONS

- A. All Selected Features
- B. Outliers Capped at \$310k
- C. Top Three Features with Highest Predictive Power

RESULTING IN 12 MODELS TO ANALYZE

WHAT DO WE HAVE?

This graph highlights the trends over time to quickly scan how the market is changing annually.

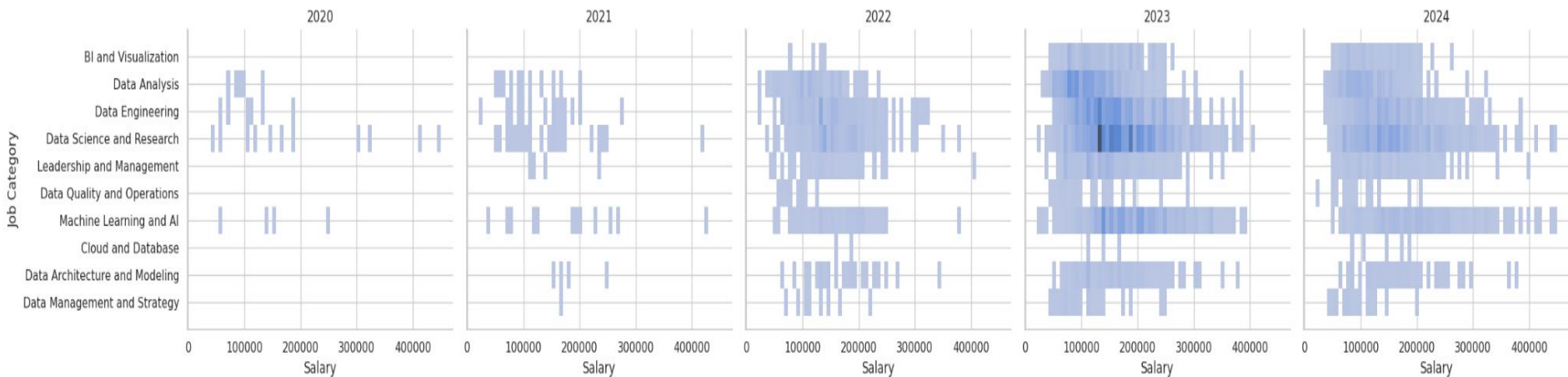
Outside of the drastic decrease in the executive level of experience, we see minimal change in average salary based on experience from year to year.



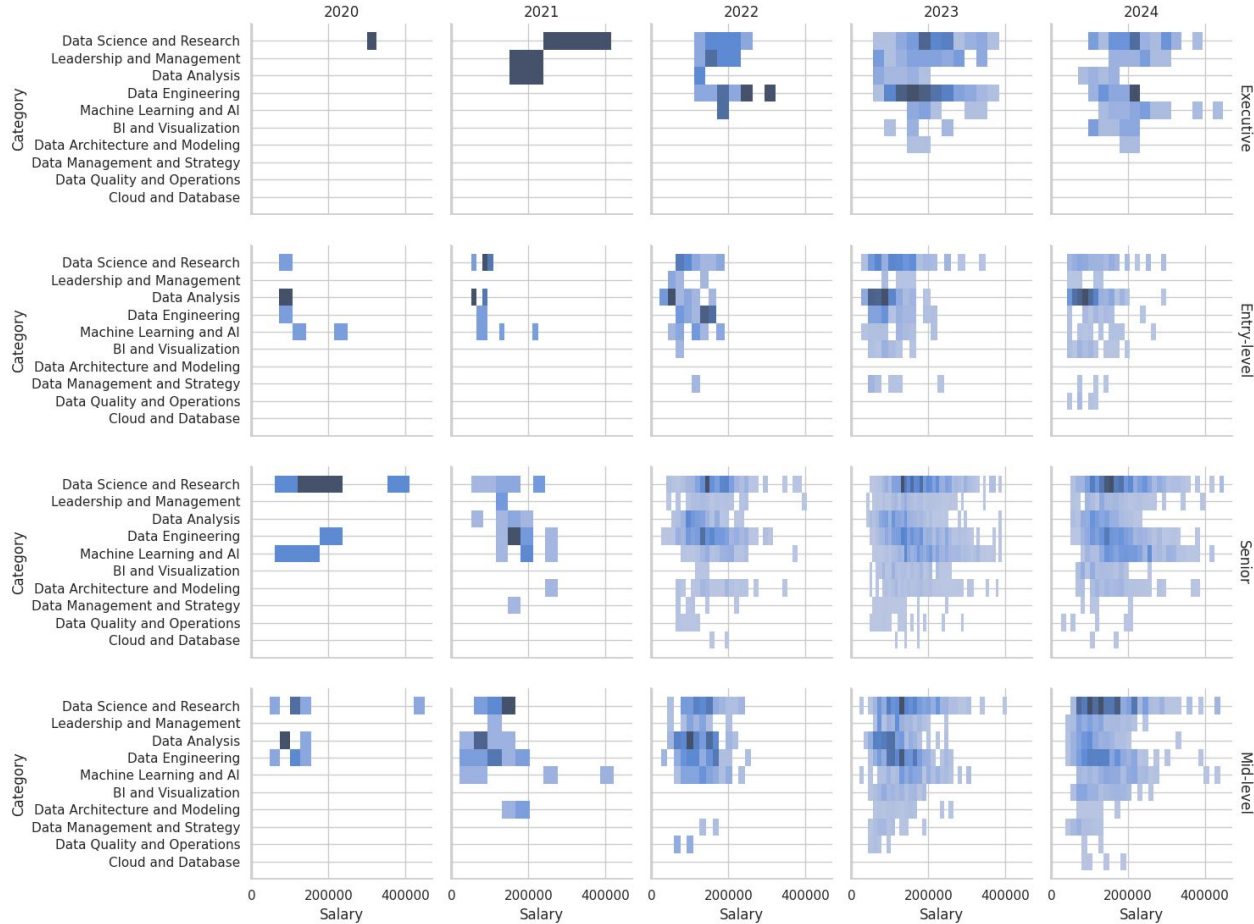
Further Exploration

Further exploration of the data set shows that while increasing each year, there is a significant difference in volume of records from 2020 compared to 2023.

Salary Records by Year and Category



Salary Information by Year, Category and Experience Level

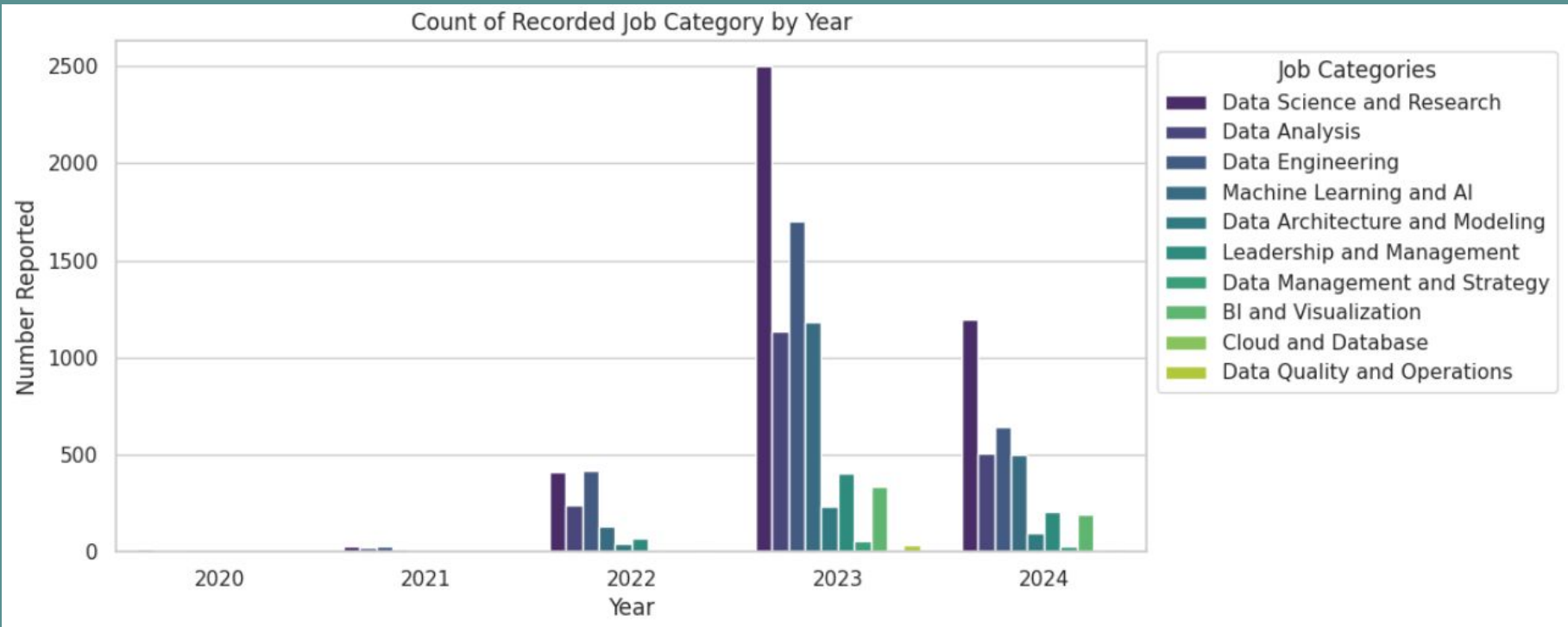


Deep Dive

This chart explores the reported salary for each job category, breaking it down by Year and Experience Level.

This visualization displays the vast change in the amount of recorded categories from year.

Trending Opportunities





Next Steps

- Additional Economic Features
- Explore Advance Feature Engineering
- Test Alternate Models
- Further Research and Hyperparameter Tuning
- Cross Validation Techniques

Conclusion

Bringing everything together, our team applied various data science methodologies to analyze and predict salary trends in today's evolving data job market. Using the data extracted from ai-jobs.net, our analysis identified key factors affecting salaries and showed the Gradient Boosting Machines (GBM) model as our top performer. However, our models fell short and did not perform as well as we expected. To address this, our plan moving forward is to enhance our dataset, explore advanced data science techniques for feature engineering, model selection, and hyperparameter tuning, and implement cross-validation to strengthen our model reliability. Ultimately, our goal is to improve our predictive capabilities to empower new graduates and professionals to make informed career and salary decisions.

References and Resources

- [A Data-Driven Approach to Salary Prediction: | by Amar saish | Medium](#): Discusses the use of Linear and Tree Regressions to predict salaries using age, experience, and other factors.
- [\(PDF\) Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review \(researchgate.net\)](#): Reviews new and existing techniques to build advanced salary predictions in Data Science, focusing on specialized skills and job benefits.
- [Employee Salaries Analysis and Prediction with Machine Learning | IEEE Conference Publication | IEEE Xplore](#): Explores various regression models to predict salaries based on influencing factors and evaluates the use of R2 and RMSE metrics.
- [Predict Data Science Salaries with Data Science | by Junting \(JT\) Lai | Towards Data Science](#): Uses EDA and machine learning algorithms to analyze the US data science job market and salaries in 2020.
- [Job Salary Prediction with NLP, Machine Learning and Deep Learning | by Bonnie Ma | Towards Data Science](#): Analyzes 10,000 plus job postings from Indeed to build a model for job salaries using machine learning and deep learning techniques.
- [Machine Learning Models for Salary Prediction Dataset using Python | Semantic Scholar](#): Highlights the result of applying three supervised machine learning techniques (linear regression, random forest, and neural networks) to a U.S. salary dataset.
- [The Recession's Impact on Analytics and Data Science \(mit.edu\)](#): Provides information regarding the potential decline in demand for data scientists and analytics due to economic disruptions.
- [Will Recession Impact Data Science and Analytics? | Analytics Insight](#): Looks into the challenges and potential declines in the demand for analytics and data science during economic recessions.
- [Key Data And Analytics Trends To Watch In 2023 \(forbes.com\)](#): Summarizes trends in data and analytics in 2023.
- [Data Scientist Job Market 2024: Analysis, Trends, Opportunities | 365 Data Science](#): Examines the current trends in the data job market, focusing on the increasing demand for data roles and the evolving job requirements heading into 2024.